Enfoque de Telefónica para un uso responsable de la IA

TELEFÓNICA, S.A.

Introducción

La inteligencia artificial (IA) está en auge. Puede aplicarse a ámbitos tan diversos como las recomendaciones de contenido, los chatbots, el reconocimiento de imágenes, la traducción automática, la detección de fraudes, los diagnósticos médicos, los vehículos autónomos, el ámbito jurídico, la educación, el transporte y la logística, por nombrar solo algunos. No solo se utiliza en los negocios, sino también para fines sociales como un mejor entendimiento y reducción de los efectos del cambio climático, los desastres naturales y la migración. Además, en Telefónica, la IA y el *big data* se utilizan cada vez más. Existen cuatro áreas principales de aplicación: i) optimización de la actividad comercial principal; ii) innovación en la relación con el cliente, utilizando tecnologías cognitivas para asistentes digitales en aplicaciones, webs, centro de atención telefónica, tiendas, etc.; iii) oferta de servicios de IA y *big data* a clientes empresariales a través del área B2B de Telefónica; y iv) y uso de la IA y el *big data* para fines sociales como la lucha contra la COVID-19 y la contribución a los Objetivos de Desarrollo Sostenible. Las dos últimas áreas se basan siempre en datos anonimizados y agregados.

Sin embargo, recientemente se ha expresado una preocupación por el uso de la IA, en particular en relación con la posible discriminación (sesgo, discriminación, paridad predictiva), la falta de interpretabilidad de las conclusiones algorítmicas (explicabilidad, problema de la caja negra) y la falta de transparencia de los datos personales utilizados.

Para hacer frente a esos posibles problemas, Telefónica publicó sus Principios de IA en octubre de 2018. Este documento describe el uso responsable de la IA por parte de Telefónica, el cual forma parte de una metodología más amplia centrada en la responsabilidad por diseño. Dado que todavía la industria tiene poca experiencia, adaptaremos y ajustaremos la metodología en función de nuestra experiencia y de los nuevos desarrollos externos.

El enfoque de Telefónica para un uso responsable de la IA

La propuesta para un uso responsable de la IA incluye:

- Un modelo estratégico: la visión estratégica sobre cómo el uso responsable de la IA está en consonancia con objetivos más ambiciosos de la empresa.
- Un modelo organizativo y de relación que define los roles necesarios, de acuerdo con la estructura corporativa, además de las funciones, las responsabilidades y las relaciones entre ellos.
- Un modelo operativo que define los principales procedimientos junto con las funciones de los responsables encargados de las tareas a realizar. Esto incluye la metodología de la IA responsable por diseño.

El modelo estratégico: principios de inteligencia artificial de Telefónica

Telefónica tiene un compromiso firme con los derechos humanos, tal y como se indica en sus Principios de negocio responsable y en la Política de derechos humanos. Estamos comprometidos con el desarrollo de productos y servicios destinados a hacer del mundo un lugar mejor para vivir o mitigar los impactos negativos que la tecnología pueda tener en la sociedad o el medioambiente. La tecnología debe contribuir a crear

una sociedad más inclusiva y ofrecer mejores oportunidades para todos, y la IA puede contribuir a estos objetivos.

Con el fin de guiar a la empresa en su aplicación de la IA y el *big data* en todas las líneas de negocio, Telefónica ha publicado sus «Principios de IA». Los principios incluyen:

- **IA justa**: Buscamos asegurarnos de que las aplicaciones de tecnología de IA produzcan resultados justos. Esto significa que no deben conducir a impactos injustos discriminatorios por razón de la raza, el origen étnico, la religión, el sexo, la orientación sexual, la discapacidad o cualquier otra condición de las personas. Aplicaremos la tecnología para reducir al mínimo la probabilidad de que los conjuntos de datos de formación que usemos creen o refuercen un sesgo o una discriminación injusta.
  Al optimizar un algoritmo de aprendizaje automático para obtener precisión en términos de falsos positivos y negativos, consideraremos el impacto del algoritmo en el dominio específico.

- **IA transparente y explicable**: Seremos explícitos con el tipo de datos, personales o no, que utilizan los sistemas de IA, así como la finalidad para la que se van utilizar estos datos. Cuando las personas interactúen directamente con un sistema de IA, seremos transparentes con ello.
  Cuando los sistemas de IA tomen o respalden decisiones, tomaremos las medidas técnicas y organizativas necesarias para garantizar un nivel de comprensión adecuado para el área de aplicación. En cualquier caso, si las decisiones afectaran significativamente a la vida de las personas, nos aseguraremos de que entendemos la lógica que hay detrás de las conclusiones. Esto también será de aplicación cuando utilicemos tecnología de terceros

- **IA centrada en las personas**: La IA debería estar al servicio de la sociedad y generar beneficios tangibles para las personas. Telefónica es consciente del hecho de que la implementación de la IA en nuestros productos y servicios no debe, en ningún caso, provocar un impacto negativo en los derechos humanos o en el logro de los Objetivos de Desarrollo Sostenible de la ONU

- **Privacidad y seguridad por diseño**: al crear sistemas de IA que están alimentados por datos, la privacidad y los aspectos de seguridad son una parte inherente del ciclo de vida del sistema. Esto garantiza al máximo el respeto al derecho de las personas a la privacidad y a sus datos personales. Obsérvese que los datos utilizados en los sistemas de IA pueden ser personales o anónimos/agregados. Obsérvese también que este principio se puede aplicar de manera más general, no solo a los sistemas de inteligencia artificial, y que Telefónica ya cuenta con procesos para garantizar la privacidad y la seguridad adecuadas.

- **Trabajar con socios y terceros**: Telefónica se compromete a verificar la lógica y los datos utilizados por los proveedores para garantizar que se respetan estos principios.

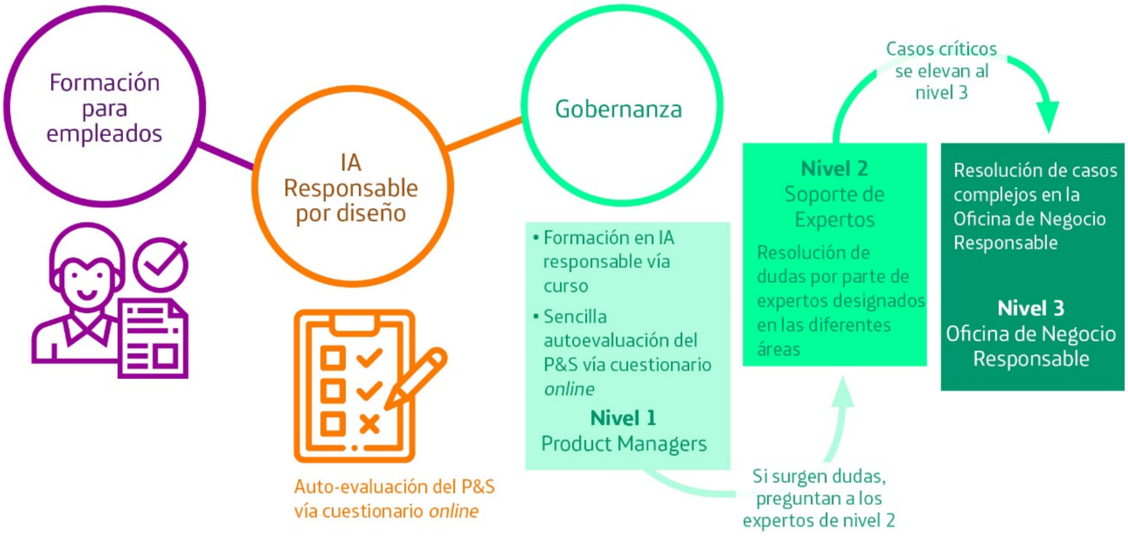
Estos principios también se basan en un amplio consenso entre las comunidades de expertos, así como en aspectos específicos de la industria de las telecomunicaciones.

Sin embargo, existe poca experiencia publicada sobre la forma en que estos principios pueden aplicarse en la práctica en las grandes empresas para conseguir el efecto deseado. En este sentido, Telefónica está dando un paso importante con este enfoque para un uso responsable de la inteligencia artificial.

El modelo organizativo y de relaciones

Estamos implementando una IA responsable a través de un modelo organizativo y de relación que define qué departamentos de la empresa se ven involucrados, cuáles son sus funciones y cómo se relacionan entre sí para alcanzar un uso responsable de la IA.

Promovemos un enfoque de auto-responsabilidad con un modelo de escalación a demanda. Existe un proceso de escalabilidad de 3 pasos que se ilustra en el siguiente gráfico.



Los jefes/desarrolladores de producto que compran, desarrollan o utilizan la inteligencia artificial deben realizar una simple autoevaluación desde el diseño del producto/servicio que están desarrollando mediante un cuestionario *online*. Esta autoevaluación trata explícitamente los posibles riesgos que puede haber para los derechos humanos relacionados con el uso de la inteligencia artificial. Esta autoevaluación se integra en un modelo de gobernanza de tres niveles apoyado por una comunidad de expertos más amplia (entre ellos, una única persona de contacto para las cuestiones relacionadas con la IA y la ética, un «IA Champion» responsable o RAI). Si un jefe/desarrollador de producto (nivel 1) tiene dudas sobre un posible impacto adverso de un determinado producto o servicio una vez completada la autoevaluación, y esta duda no puede resolverse con la ayuda de un RAI, se le planteará estas dudas automáticamente a un grupo de expertos multidisciplinar de la compañía (nivel 2), que junto con el jefe/desarrollador de producto intentarán resolver el problema. En caso de que sea un riesgo potencial para la reputación de la empresa, el asunto se elevará a la Oficina de Negocio Responsable, que reúne a todos los directores de departamentos relevantes a nivel global (nivel 3).

El modelo operativo

El modelo operativo describe los procedimiento para implementar el enfoque de IA responsable en el día a día de la empresa. Integrada dentro de una visión más amplia de la responsabilidad por diseño, incluye una metodología llamada «IA responsable por

diseño» inspirada en metodologías ya existentes en privacidad y la seguridad por diseño. El modelo operativo consiste, entre otras cosas, en:

- Actividades de formación y concienciación
- El cuestionario de autoevaluación online, donde cada principio de la IA se pone en práctica a través de una serie de preguntas para responder y un serie de recomendaciones.
- Un conjunto de herramientas técnicas que ayudan a responder a las preguntas

Para proceso de diseño de la metodología se requirió de una iniciativa interempresarial en la que participaron diferentes departamentos como el de Ingeniería, Ética y Sostenibilidad Corporativa, Seguridad, Jurídico, Comercial, Recursos Humanos, Compras, así como el respaldo del equipo directivo.

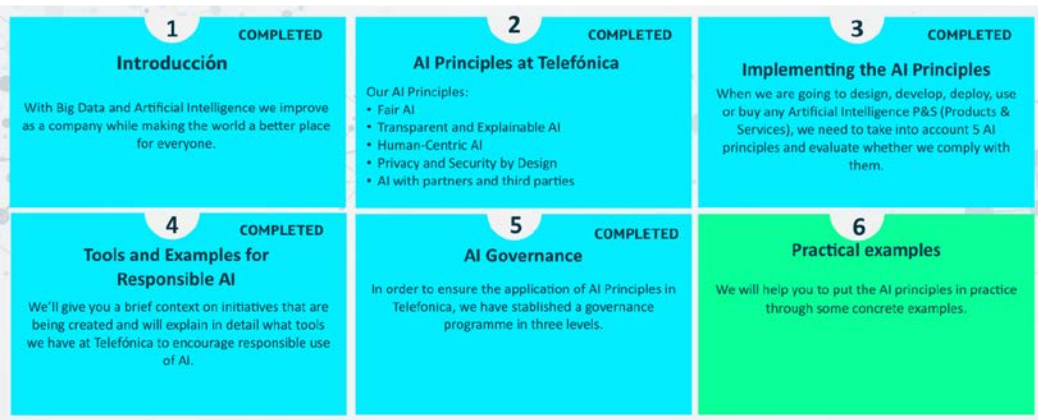
Formación y concienciación

Si bien la inteligencia artificial es todavía una tecnología bastante nueva que se utiliza en las grandes empresas, sus repercusiones éticas y sociales son todavía más recientes. Por lo tanto, es de suma importancia explicar a los trabajadores lo que es la IA, cómo funciona y cómo puede derivar en consecuencias no deseadas.

Por ello, Telefónica ha desarrollado unos cursos relacionados con la IA y la ética que son accesibles a todos los empleados a través de los portales corporativos habilitados y en tres idiomas (español, inglés y portugués). Dependiendo del perfil del empleado, puede acceder a una versión más corta del curso, que se divide en tres módulos de 10 minutos cada uno, o a una versión más extensa, que dura aproximadamente una hora.



El curso completo está dividido en 6 módulos, tal y como se muestra en la siguiente imagen:



Aparte de los cursos *online*, cada unidad de negocio que utiliza la IA participa en talleres dedicados a explicar el modelo de gobernanza. Además, existe una miniguía disponible en la intranet donde los empleados pueden obtener una visión general de cómo aplicar los principios de la IA para garantizar un uso responsable de la misma.

Cuestionario con recomendaciones

Para todos los productos y servicios que utilizan IA o *big data*, el jefe de producto debe completar el cuestionario de autoevaluación en el que se plantean varias preguntas para cada principio. El cuestionario está disponible *online* en español e inglés y se integra dentro de la iniciativa global «Responsabilidad por diseño» del grupo Telefónica. Todos los cuestionarios completados se registran para su análisis, estadísticas y acciones si así se requiere. Los responsables que hayan rellenado el cuestionario recibirán un correo electrónico con su cuestionario completo, un conjunto de recomendaciones si procede y una indicación de las cuestiones que deberán resolver o volver a revisar más adelante.

El proceso de diseño de la metodología requirió una iniciativa interempresarial en la que participaron diferentes departamentos como el de Ingeniería, Ética y Sostenibilidad Corporativa, Seguridad, Jurídico, Comercial, Recursos Humanos, Compras, así como el respaldo del equipo directivo.



El cuestionario consta de 12 preguntas que abarcan todos los principios de la IA. En la imagen que aparece a continuación se puede ver un ejemplo:

PRINCIPIO 1: JUSTA

Nos aseguramos de que las aplicaciones no conducen a resultados con sesgos e impactos discriminatorios e injustos.

1. Si hay datos sensibles en el/los set de datos y su uso no está justificado para el producto o servicio que estamos desarrollando:
- Para IA: ¿Se ha comprobado durante el proceso de aprendizaje que el modelo no discrimina con respecto a estas variables sensibles?
- Para Big Data: ¿Se ha evitado usar los datos sensibles para los procesos?
Datos o variables sensibles son datos especialmente protegidos, tales como: ideología, afiliación sindical, convicciones religiosas, creencias, origen racial o étnico, datos relativos a la salud, datos genéticos, datos biométricos, orientación sexual y vida sexual.

Si	No	N/A
○	◉	○

En caso de usar aprendizaje automático: En caso de que haya variables sensibles en el data set y su uso no sea justificados por la naturaleza propia del servicio (p.e. una aplicación para diagnostico o monitorización medica), hay que borrarlos durante el proceso de entrenamiento del algoritmo. Si quiero asegurarme de que este tipo de datos sensibles no influyen en el resultado del algoritmo, puedo incluirlos a posteriori (después del entrenamiento) con el único propósito de verificar que el modelo no discrimine según estos variables sensibles.
En caso de usar big data (pero no aprendizaje automático): En caso de que haya variables sensibles en los datos y su uso no sea justificados por la naturaleza propia del servicio (p.e. una aplicación para diagnostico o monitorización médica), hay que borrar estos datos sensibles.

Es importante destacar que, aunque la privacidad y la seguridad forman parte de los «Principios de IA» de Telefónica, la empresa cuenta con procesos y áreas específicas dedicadas a la privacidad y la seguridad. De hecho, la privacidad y la seguridad son fundamentales para cualquier sistema digital y no solo para las aplicaciones de IA y *big data*. Por lo tanto, en cuanto al principio de privacidad y seguridad, el cuestionario se refiere a las respectivas áreas de la empresa: al delegado de protección de datos (DPD) y al responsable de seguridad interna (CISO).

Herramientas técnicas

Dado que algunas de las preguntas del cuestionario son imposibles de responder sin herramientas específicas, nuestra metodología incluye tanto herramientas internas como externas (en su mayoría de código abierto). Algunas de las herramientas internas se relacionan con tecnologías para mejorar la privacidad, como la anonimización, un instrumento para detectar la discriminación algorítmica contra grupos protegidos y un instrumento de transparencia de los datos personales. Constantemente estamos evaluando nuevas herramientas externas, sobre todo las relacionadas con el sesgo, la discriminación y la explicabilidad. Se trata de un ámbito de investigación muy activo en el que aparecen nuevas herramientas a un ritmo rápido. La comunidad de expertos es la responsable de analizar y recomendar la inclusión de nuevas herramientas.

Otras referencias de Telefónica

Towards organizational guidelines for the responsible use of AI, Richard Benjamins, 24th European Conference on Artificial Intelligence (ECAI), 2020, Santiago de Compostela, España, https://ecai2020.eu/papers/1347_paper.pdf

Responsible AI by Design in Practice, Richard Benjamins, Alberto Barbado, Daniel Sierra, Actas de la Human-Centered AI: Trustworthiness of AI Models & Data (HAI) en el AAAI Fall Symposium, Washington DC, del 7 al 9 de noviembre de 2019, https://arxiv.org/html/2001.05375, https://arxiv.org/abs/1909.12838

Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI, Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, Raja Chatila, Francisco Herrera, Information Fusion, volumen 58, junio de 2020, páginas 82-115.

Towards a framework for understanding societal and ethical implications of Artificial Intelligence, Richard Benjamins, Idoia Salazar, en *Vulnerabilidad y cultura digital. Riesgos y oportunidades de la sociedad hiperconectada*, Dykinson, páginas: 2020, https://www.dykinson.com/libros/vulnerabilidad-y-cultura-digital-riesgos-y-oportunidades-de-la-sociedad-hiperconectada/9788413246475/